Beyond Traditional Dialects: The Case for Diaglossic Modeling of European Linguistic Variation

Anonymous ACL submission

Abstract

Modern speech processing models struggle with dialectal variation, primarily due to the limited amount of speech data for non-standard varieties. In this position paper, we examine the European linguistic landscape and the coarse-grained view on its dialectal variation that modern language technology adopts. Limited to horizontal modeling of diatopic variation (i.e., gradual change of language with geography), current approaches, we argue, neglect the important *diaglossic* dimension: the vertical continuum between the traditional dialects and the high standard of a language. Our laborious examination of 86 speech datasets highlights the critical gap between the diatopic and diaglossic annotation: while many efforts collect speech from localized areas (e.g., regions or villages) to capture "dialectal" speech, the instances are often much closer to the high standard than traditional dialects. Further, we point to complex interactions where diaglossic continua of standardized languages overlap with one another and/or varieties of weakly standardized languages. Finally, we argue that supporting non-standard varieties requires a more effective *cross-variety transfer*, which hinges on both a meaningful modeling of diaglossic continua and interactions between varieties. To this end, we recommend (1) creation of computer-readable (a) *diaglossic hierarchies* for standardized European languages, and (b) *interaction registers*, documenting interactions between varieties of different languages, and, crucially, (2) annotating dialectal speech instances with diaglossic information, primarily the diaglossic distance to the high standard.

1 Introduction and Motivation

Modern language technology (LT) meaningfully supports only a fraction of the world's 7,000+ languages—a fact that every effort to broaden the multilingual scope of LT seems to mention, whether on (i) text-based language modeling (NLLB Team et al., 2022; Lai et al., 2024, *inter alia*), (ii) automatic speech recognition and speech representation (Chen et al., 2024; Omnilingual ASR team et al., 2025, *inter alia*), or (iii) evaluation benchmarks (Conneau et al., 2022; Ahuja et al., 2024; Schmidt et al., 2025; Keller and Glavaš, 2025, *inter alia*). Though there clearly exist massive and systemic disparities in LT performance across standardized/institutionalized languages (Blasi et al., 2022; Alhanai et al., 2025), there is another dimension of inequity in LT support that receives much less attention (Faisal et al., 2024): that between standardized and non-standard (i.e., dialectal) varieties of languages. The latter often lack a canonical written norm and are often predominantly used in spoken contexts (Blaschke et al., 2023a; Ramponi, 2024; Joshi et al., 2025).

The causes of this gap are twofold. First, from a computational point of view, training data for supporting non-standard varieties are typically orders of magnitude smaller than for the respective standardized varieties, making data-driven language learning inherently difficult (Kantharuban et al., 2023). Second, from a sociolinguistic perspective, most speakers are also fluent in the standard variety of their (or other co-existing) language(s) (Auer, 2017; Lutgen et al., 2026), which, in the European context, mitigates the effect of digital isolation: supporting varieties of well-supported standardized languages is thus perceived less pressing than supporting institutionalized languages with weak support. From the technical perspective, most methodological developments target standard languages (Plank, 2016; Lutgen et al., 2026); dedicated efforts to develop LT for non-standard varieties are scarce (Aepli and Sennrich, 2022; Blaschke et al., 2023b; Hofmann et al., 2024; Feng et al., 2026) and even those target coarse-grained varieties such as regional standards (e.g., the Bosnian variety of BCMS) or regional variants, i.e., weakly defined semi-standardized varieties between the traditional

dialects and the standard language (e.g., Bavarian as a "dialect" of German). Efforts that investigate LT support for dialectal variation at a finer level of granularity are instead very rare (Chifu et al., 2024; Souganidis et al., 2026). Critically, local and minority languages—each with its own dialects—are often framed as monolithic, single-dialect entities with an established written form, suggesting that the inner variation and primarily-oral use of such languages are largely overlooked (Ramponi, 2024).

In this position paper, we challenge this coarse-grained view of language variation that current LT adopts and argue that analyses of LT support for non-standard varieties need to consider two orthogonal continua: the usually considered (1) *horizontal* (*diatopic*) continuum (Trudgill, 1974; Jeszenszky and Weibel, 2015; Hofmann et al., 2024), i.e., a gradual change of language with geography needs to be coupled with the (2) *vertical* (*diaglossic*) continuum, that spans varieties in between the traditional dialects (subject to horizontal variation) at the bottom and the high standard at the top, with intermediate varieties standardizing monotonically increasing subsets of linguistic features. In contrast to existing work that surveys text-based dialectal LT (Mashabi et al., 2024; Joshi et al., 2025), we focus on fine-grained modeling of spoken-language variation, with a focus on European languages.

The case for European dialects. We focus on spoken varieties of European languages[1] for three main reasons. First, most European languages are high-resource, fully standardized languages with strong LT support,[2] providing meaningful reference points for non-standard varieties; in other parts of the world (e.g., Africa, Southeast Asia) LT support is often limited even for standard/institutionalized varieties (Wu et al., 2025). Second, as a consequence of longstanding and iterative standardization pressure, most European languages exhibit diaglossic variation (see §2.1), i.e., there is a non-negligible vertical continuum between the traditional dialects and the standard language (Auer, 2011). Third, we (i.e., all authors) are European researchers with experience in studying the variation of European languages. By excluding other geographies, we not only stay within our area of expertise, but also reduce the risk of extractive research and cultural appropriation (Ògúnremí et al., 2023; Bird, 2024).

The case for spoken language. Human languages are primarily spoken and then, sometimes, written (Chrupała, 2023).[3] The need to support dialectal speech—via automatic speech recognition (ASR) and/or spoken language understanding (SLU) models—is much more pronounced than the need to support dialectal text for several reasons. First, most European languages are institutionalized, and the writing of the standard variety is taught in schools, which means that: (1) most dialectal speakers are proficient in a written standard variety, eliminating the need to write the dialect,[4] which in turn leads to (2) most dialects not having a standardized orthography. Second, the European linguistic space is largely diaglossic. In the continuum between base dialects and high standards, phonetic/phonological features have been shown to remain dialectal longer than morphosyntactic features (Trudgill, 1986); this renders speech a modality with stronger dialectal differentiation. Finally, there is recent evidence that non-standard language speakers care more about dialectal speech understanding (i.e., as input) than dialectal output (Blaschke et al., 2024).

Besides extending the scope of multilingual LT to ever more (standardized) languages, we argue that truly inclusive LT also has to account for dialectal speech *across both horizontal and vertical axes of variation*, from regional varieties to base dialects. We believe that this should be one of the key new missions for LT research.

2 The Complex Continua of European Language Variation

Dialectal variation is commonly observed through a set of interconnected, fluid, and sometimes overlapping continua, over which linguistic features (e.g., phonetics, syntax, vocabulary) gradually change. In sociolinguistics, these continua are jointly called *diasystem* (Weinreich, 1954; Höder, 2014; Gordon, 2017). The following are commonly introduced as the key continua (de Saussure, 1916; Coseriu, 1956; Mioni, 1983): (i) *diatopic*: linguistic differentiation over physical distances (i.e., geographic varia-

[1] We make no distinction between European languages based on their political status (e.g., official EU language or not, or recognition at the country or regional levels only). We only exclude English due to (i) its geographic dispersion and (ii) extensive coverage of spoken English variation in other work, e.g., SPADE (Stuart-Smith et al., 2024).

[2] E.g., 56 out of 100 largest Wikipedias (incl. all 24 official EU languages) by article count are in European languages.

[3] According to Ethnologue, more than 3,000 languages lack a commonly used, developed, or standardized orthography.

[4] A dialect can belong to an institutionalized language or a local language with no established standard (see §2.2).

tion); arguably the dimension of linguistic variation that first comes to mind when we talk of dialects; (ii) *diachronic*: continuum that reflects linguistic changes over time; (iii) *diastratic*: variation due to social stratification (e.g., education level, social class, or ethnicity); (iv) *diaphasic*: differentiation due to the situation or context (e.g., formal vs. informal); and (v) *diamesic*: variation caused by the medium of communication, and in particular the differences between the spoken and written forms. While a complete computational treatment of linguistic variation warrants considering all these continua, modern data-driven LT, however, has to keep in mind the practicality of the data collection for their computationally modeling. Because of this, most work on dialectal LT targets (coarse-grained) diatopic (Zampieri et al., 2020; Blaschke et al., 2023a; Ramponi, 2024; Joshi et al., 2025; Feng et al., 2026) or diachronic variation (Kutuzov et al., 2018; Loureiro et al., 2022; Fittschen et al., 2026).

In this position paper, we make a case for modeling of a *diaglossic* continuum of language variation, which has, computationally, largely been neglected. Especially in Europe, where standardization efforts for many languages span centuries, the computational modeling of (1) the *diatopic* (or, *horizontal* or *geographic*) continuum, i.e., how linguistic features change over geography, needs to be coupled with the modeling of (2) *diaglossic* (or, *vertical* or *standardization*) continua (Auer, 2011), where intermediate varieties that standardize increasingly large subsets of linguistic features bridge the gap between base dialects and the standard variety.

2.1 The Diaglossic Continuum

The diaglossic dimension of dialectal analysis reflects gradual linguistic *standardization*: the vertical distance of a variety from the base—the surface geography, which exhibits diatopic variation—indicates the degree of standardization of that variety with respect to the corresponding base dialects (Figure 1). At the "roof" of a diaglossic continuum is the most standardized variety, typically with a "high-prestige" function (Kloss, 1978; Auer, 2011), seen as providing a roof for all vernaculars of a language. Auer (2011) defines three properties of an ideal *standard* variety over a geography with diatopic variation: (i) *invariance*: this variety, ideally, exhibits no geographical variation; (ii) *prestige*: the variety has a clear prestige over vernaculars, and is a variety of preference in formal settings; and (iii) *codification*: expressions and utterances can

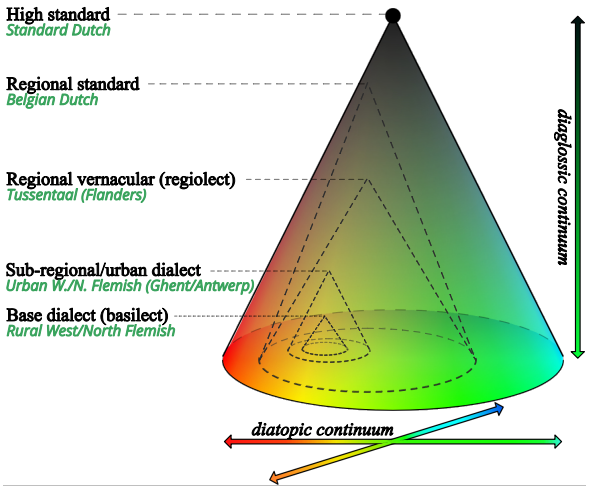

Figure 1: Diatopic and diaglossic language variation, with one possible discretization of the diaglossic continuum: from base dialects (basilects) over sub-regional and regional vernaculars (regiolects) to regional standards and the high-standard variety.

be judged for correctness (regardless of whether there exist formal codifications in the form of grammar books, dictionaries, etc.). We deem prestige largely inconsequential for data-driven modeling; similarly, while the impact of codification (e.g., the existence of grammar books and dictionaries) is significant for traditional computational linguistics, it arguably diminishes under data-driven language modeling. In what follows, we thus focus on geographic invariance as a property of interest for data-driven modeling of diaglossic variation.

Formalization. We first provide a formalization of the diaglossic continua, as follows. Let F_D denote the set of all linguistic features, e.g., typological features (phonology, morphology, syntax) such as those documented in resources like WALS (Dryer and Haspelmath, 2013) or Phoible (Moran and McCloy, 2019) or lexical features, that exhibit *some degree of variation* across some diatopic continuum (i.e., geography) D. We formalize a diaglossic continuum $\mathcal{D} = (D, \succ_{dg})$ as a partially ordered set—a set of language varieties spoken across the diatopic continuum, $D = \{d_1, d_2, \ldots, d_N\}$, partially ordered by a (transitive and antisymmetric) relation of diaglossic subsumption $\succ_{dg}$. We associate with each variant d_i a set of linguistic features $F_i \subseteq F_D$ that exhibit variance over the underlying geography D_i (base of the respective diaglossic cone, see Figure 1) and which are made invariant (i.e., standardized) in the variant d_i. A variant d_i *diaglossically subsumes* (or, is *diaglossically higher* than) a variant d_j, i.e., $d_i \succ_{dg} d_j$ if and only if $F_i \supset F_j$.

At the bottom of a diaglossic continuum are the base dialects or *basilects*: they have no diaglossic height, i.e., they cover the smallest areas across which all features are geographically invariant. In other words, if d_i is a basilect, then $F_i = \varnothing$. At the opposite end of a continuum is the most standardized variety of a roof language, commonly called *high standard*. Ideally, a high standard makes all linguistic features invariant across the entire D, i.e., if d_i is a high standard, then $F_i = F_D$. In between basilects and a high standard, there is a continuum of variants that standardize an increasing subset of features over the underlying geographies. Note that, in principle, a variant may be directly (i.e., not just transitively) diaglossically subsumed by more than a single variant, i.e., $d_i \succ_{dg} d_k$ and $d_j \succ_{dg} d_k$, with no diaglossic subsumption between d_i and d_j.[5] Because of this, diaglossic continua must be modeled as directed acyclic graphs and not trees.

Unifying discretization. The inherent impracticality of determining heights of diaglossic varieties leads to the introduction of generic discrete diaglossic levels (Auer, 2005; Schmidt, 2011):

Basilects are local varieties that are farthest from the standard language and closest to traditional local speech, associated with (i) strong local grammatical, vocabulary, and/or pronunciation features, (ii) rural areas and older speakers, and (iii) intelligibility difficulties for outsiders.

Urban or sub-regional vernaculars are less localized than basilects and considered a result of urbanization and localized migration. These varieties generally "smoothen" the very local (i.e., village) features of respective basilects and mix some elements of more standardized varieties. Commonly instantiated in casual speech, they are associated with young and middle-aged speakers. These varieties weakly standardize pronunciation features of basilects and can still pose an intelligibility challenge for outsiders (Schmidt, 2011).

Regional dialects (regiolects) are larger regional speech varieties that are generally considered mutually intelligible with standard varieties and understandable to most outsiders from the same diatopic space (i.e., same roof language). These varieties typically retain a regional "accent" as well as some dialectal grammatical and lexical features.

[5] E.g., rural dialects around Rothenburg ob der Tauber in southern Germany are both diaglossically subsumed by the Upper East Franconian (a regional prestige variety to the east) and Swabian (a regional prestige variety to the southwest).

Regional standards are very close to the high standard; they can be seen as high-standard with "regional coloring", manifested predominantly in speech with dialectal prosody and phonology (Hove, 2013; Verhoeven, 2005). Grammar and vocabulary of a regional standard largely mirror those of the high standard. These varieties essentially serve as the high standard within the region and are used in formal contexts (e.g., media or education).

High standard (acrolect) is the high variety closest to the codified standard of a language. One can think of an acrolect as an (imperfect) instantiation of the codified standard. It is the variety designated for supraregional communication, used in formal education, national media, and administration.

While a practical necessity, these discrete labels do not necessarily indicate the same diaglossic heights (i.e., vertical distances from the acrolect). Across Europe, varieties denoted by the same discrete label, e.g., a *regiolect*, inevitably appear at different diaglossic heights within their respective continua. For example, the Urban East Norwegian is much closer to standard Bokmål than Français Méridional is to standard French, but both are considered regiolects. In fact, the same often holds for varieties within a single diaglossic continuum: local varieties in one geographic subarea may be much closer to the acrolect than varieties from another region referred to with the same discrete diaglossic label, e.g., regiolect. This is especially true for varieties on which the standard variety was based (e.g., Upper Saxon for Hochdeutsch, that is, standard German). The vertical spectrum can thus be interrupted at places: e.g., it is interrupted in Northern Germany, but continuous in Southern Germany (Kehrein, 2019); the degree of (full) diaglossic continuity and height of the continuum directly depend on the linguistic distance between the local basilect and the standard language (Kloss, 1978; Goebl, 1986). Even for the monolingual areas, our illustration of a diaglossic continuum from Figure 1 is a simplification: in reality, the base is not flat, as basilects differ in their linguistic distance to the acrolect. And the situation becomes much more complex in multilingual areas.

2.2 Multilingual Continua

So far, we only considered the simplest case: a vertical continuum of a single language (Figure 1 and Figure 2a). Many European varieties, however, co-exist and interact with varieties of other

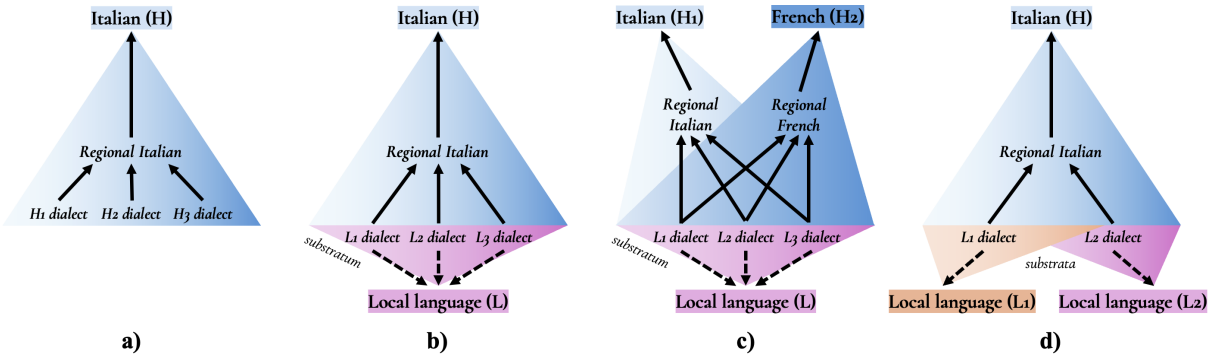

Figure 2: Illustration of dialectal complexity in Europe: **a)** Monolingual diaglossic continuum of a standardized (*high*) language H (e.g., Italian); **b)** Interplay between a standardized language H and a (closely) related non-standardized (or weakly standardized; *low*) language L; **c)** Interaction of diaglossic continua of two (or more) high languages, H_1 and H_2 (e.g., Italian and French), and a low language (e.g., Franco-Provençal), (closely) related to (at least) one of the high languages; **d)** Interplay between a high language H (e.g., Italian) and two (or more) low languages, L_1 and L_2: typically, an L (e.g., Sicilian) is (closely) related to H, whereas the other L (e.g., Italiot-Greek) is distant from H and spoken by a different ethnic community, making the dialects of the latter *ethnolects*.

languages spoken in the same area. We will distinguish between two types of languages based on their language functions (Fishman, 2001) and thus the extent of their standardization: (1) "*high*" (H) languages (e.g., Italian, Croatian) are extensively standardized and usually exhibit a full diaglossic continuum; (2) "*low*" (L) languages have little to no standardization and thus a shallow vertical continuum. To illustrate multilingual continua of varying complexity, we turn to the geographic area of modern-day Italy, which offers diversified linguistic configurations, owing to (1) the renowned linguistic history of the area and (2) the relatively late adoption of Italian as a unifying language.

Figure 2 illustrates multilingual continua in order of increasing complexity. Figure 2b shows a configuration commonly observed in Italy: the interaction between Italian (H) and dialects of a local language L developed separately from Vulgar Latin (e.g., Sicilian, Piedmontese, Ligurian, Venetian), which leads to a diaglossic continuum in the H language. The everyday language in such areas is a variety between a dialect of L (e.g., Palermitan dialect of Sicilian) and some diaglossic variety of H (e.g., Extreme Southern Italian as an Italian regiolect). While particularly abundant in Italy, such bilingual continua are found across Europe: e.g., Finnish (H) and Karelian (L), Modern Standard Greek (H) and Tsakonian (L), or Romanian (H) and Aromanian (L). While the Iberian Peninsula may seem to offer similar cases with Spanish vs. Catalan or Galician, the latter have been standardized and cannot be considered L languages: the interactions thus occur between two related high languages. Figure 2c illustrates overlapping diaglossic continua of two co-existing high languages, H_1 and H_2 (e.g., Italian and French), but in a more complex configuration: here, these two continua additionally interact with Franco-Provençal as an L language; a dialect of L (e.g., Courmayeur Patois) here contributes to form diaglossic varieties of both H_1 (Northern Italian) and H_2 (Aostan French).

Minority languages and ethnolects. Finally, Figure 2d illustrates multilingual continua that include languages brought by an external ethnic group via migration. The "imported" language is typically not closely related to the local language(s) and, through interaction with them, becomes a type of *ethnolect*. One example is Italiot-Greek spoken in southern Italy. Its Griko dialects, spoken in nine villages in the Salento area in Italy, co-exist with local dialects of the Extreme Southern language group (often referred to as Sicilian; L) as substrata and jointly contribute to an Extreme Southern Italian (H) regiolect. Other examples of such configurations in Italy include Arbëresh (derived from old Tosk Albanian) and Molise Croatian (developed from the Ikavian Chakavian dialect of Serbo-Croatian) as ethnolects. In other geographies, the makeup is usually simpler: an ethnolect interacts with one H language (e.g., Karaševski Croatian in Romania, Donauschwäbisch in Bosnia and Serbia, or Ceylon Portuguese in the Netherlands).

Though ethnolects are sometimes seen as dialects rather than separate languages (e.g., Italiot-Greek is typically considered a dialect of Modern Greek), we argue that—from the diaglossic perspective of the destination geography—they have

to be treated as local L varieties. This is because migration carves them out of their original diaglossic continua (if one even existed at the time of migration). Their speakers only know and use that local variety (e.g., Karaševski Croatian), dissociated from the respective H language (i.e, BCMS, formerly known as Serbo-Croatian) with a notable temporal distance; they do not know or use any other variety from the H's diaglossic continuum. These minority varieties borrow features from the local languages (e.g., Italian and Sicilian for Griko), while conserving features that have disappeared in related varieties in the source geography. The variety thus becomes an *abstand language*, i.e., sufficiently distant from the roof to be considered a separate language (Kloss, 1967; Trudgill, 1992).

3 Towards Fine-Grained Computational Modeling of Complex Speech Continua

Numerous individual efforts, as well as projects and initiatives[6] collect (multilingual) European speech data, including for many dialectal varieties (Komrsková et al., 2017; Barcala et al., 2018; Wang et al., 2021; Blaschke et al., 2025, *inter alia*). Despite this, the dialectal speech corpora remain limited in size, more appropriate for evaluation than meaningful training of ASR or SLU models.

3.1 The Dialectal Speech Technology Gap

We first illustrate the gap in performance that ASR models exhibit between standard H varieties and respective low-resource dialectal varieties: basilects, ethnolects, and (closely) related L languages. We compare ASR performance for (1) the Bizkaian dialect of Basque (single basilect) (Castelruiz et al., 2004) vs. "standard" Basque from Common Voice; (2) Griko (ethnolect) (Zanon Boito et al., 2018) vs. Common Voice Greek; (3) the Metropolitan Naples variety of Neapolitan (closely related L language) (Cacioli et al., 2026) vs. Common Voice Italian; and (4) a collection of rural Spanish varieties (multiple basilects) (COSER dataset; de Benito Moreno et al., 2016) vs. Common Voice Spanish.[7] We measure the ASR performance of three state-of-the-art multilingual ASR models: Whisper (*Large-v3*) (Radford et al., 2022) and two model variants (CTC- and LLM-based) of Omnilingual-ASR (Omnilingual ASR team et al., 2025). The results, shown in Figure 3, paint a dire picture. For each dialectal variety, all models perform drastically worse than for the respective (more) standard speech. In fact, there is no model-dialect combination that produces a Character Error Rate (CER) below 25%, the usual utility threshold for ASR. This suggests that the differences between the standard variety and the L dialects are too large for successful transfer learning. We can think of only two remedies: (i) large-scale speech collection for each L variety and/or (ii) more effective transfer, from diatopically and, crucially, diaglossically closer variants.

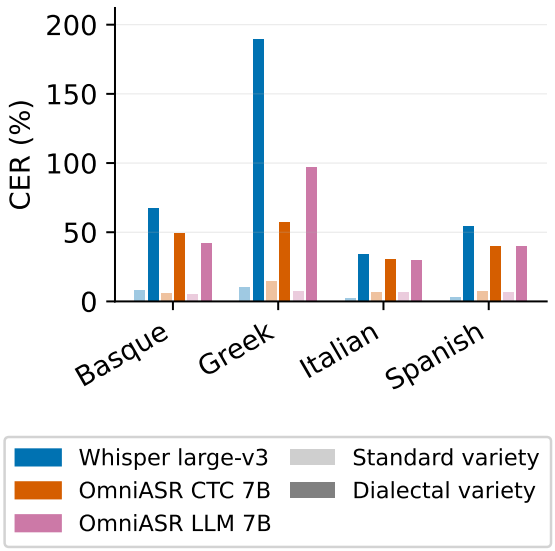

Figure 3: Performance of multilingual ASR models on pairs of "standard" (i.e., diaglossically high) and dialectal varieties (basilects, ethnolects, and closely related L languages) across four European languages. Metric: Character Error Rate (CER; lower is better).

3.2 Facilitating Cross-Variety Transfer

While it is infeasible to collect large quantities of speech for thousands of European dialectal varieties, a more effective cross-variety transfer of speech models to such varieties may be possible. However, based on our analysis of the complexity of European linguistic continua from §2, unlocking it warrants (a) detailed diaglossic hierarchies of H languages; (b) a register of interaction languages, primarily for L varieties and, crucially, (c) speech instances annotated for diaglossic height.

Genealogical vs. diaglossic hierarchies. Popular large-scale linguistic resources like Glottolog (Hammarström et al., 2026) organize languages *and dialects* into genealogical hierarchies. These

[6] E.g., CLARIN Spoken Corpora and Common Voice.

[7] Note that, while the Common Voice portions of respective H languages do not contain (only) the high-standard speech, the average diaglossic height of their instances is in any case well above the basilect data against which we compare.

are not really useful for facilitating cross-variety transfer for three key reasons. First, they reflect historical, evolutionary relations between dialects, rather than their contemporary linguistic similarities; intermediate varieties in these hierarchies are often dead historical varieties. Second, for many parts of Europe, they are too coarse-grained for practical applications. Third, and most critically, these hierarchies are largely oblivious to modern-day country boundaries and standardization effects that result in overlapping diaglossic varieties.

Modern Croatian dialects illustrate the inadequacy of genealogical hierarchies well. Glottolog distinguishes between Kajkavian and Serbian-Bosnian-Croatian (BCMS), with BCMS as the parent of Chakavski and Shtokavski. Modern-day Croatia, however, encompasses Kajkavian, Shtokavian, and Chakavian dialects, all three of which have diaglossically high(er) varieties induced by modern standard Croatian. This brings the diaglossically higher Kajkavian varieties much closer to Shtokavian dialects (the core of the standard Croatian), which represents a practically important relation that Glottolog simply does not have.

Facilitating cross-variety transfer, therefore, calls for a construction of diaglossically-informed dialectal hierarchies, driven by the relation of diaglossic subsumption $\succ_{dg}$ (see §2.1). To this effect, modern-day dialectal classifications, such as those introduced for many European languages within the DIACLEU initiative (Aurrekoetxea et al., 2022), offer a better starting point than dialectal genealogy. Unfortunately, most such classifications are provided only descriptively (e.g., in scholarly publications), requiring a non-trivial conversion into computer-readable formats. These classifications still need to be carefully examined through the diaglossic lens and appropriately adjusted, because the mapping between dialects in existing classifications and truly spoken diaglossic varieties is still not one-to-one. We identify two common types of misalignment: **(1)** intermediate dialects from existing classifications do not have a corresponding diaglossic variety; e.g., while a diaglossic variety "Istrian Croatian" largely corresponds to the Southwestern Istrian sub-dialect of Chakavian, there is no real diaglossic variety that corresponds to Chakavian itself; conversely, **(2)** there are real diaglossic varieties without a corresponding dialect in existing classifications; for example, while Southern, and Southeastern Chakavian, have corresponding diaglossic levels, a strong case can be made for a higher diaglossic level of "Dalmatian Croatian", not reflected by the existing dialectal classification.

Dialect	Language	Ethnolect?	Interactions
Palermitan	Sicilian (L)	✗	Extreme Southern (Italian, H)
Courmayeur patois	Franco-Provençal (L)	✗	Northern Italian (Italian, H) Aostan French (French, H)
Calimera Griko	Italiot-Greek (L)	✓	Extreme Southern (Italian, H) Central Salentino (Sicilian, L)

Table 1: Example of interaction entries for varieties of L languages (i.e., without developed diaglossic continua).

Defining the hierarchy of diaglossic variants for European H languages is just the first step. We argue that each variant additionally needs to be labeled with a *diaglossic height* (or depth), a quantity that indicates its linguistic distance to the high standard. As emphasized in §2.1, commonly used diaglossic labels (e.g., basilect, regiolect) are vague and unreliable indicators of a variety's diaglossic height. For languages whose dialectal varieties have detailed annotations of linguistic features F_D (see §2.1), e.g., Dutch (Barbiers et al., 2007; Goeman and Taeldeman, 1996) or German (Limper et al., 2019), one can derive diaglossic heights from distances between the varieties' feature vectors and the vector of the high standard. We do acknowledge that most varieties of European languages do not have such linguistic descriptions readily available and that annotating hundreds of linguistic features for all varieties is infeasible. But we believe that these are not strictly necessary: even reasonable estimates of vertical distances of dialectal varieties to the high standard—which linguists specialized in a particular H language can certainly provide—will suffice for a more informed and effective cross-variety transfer of speech models.

Interactions register. We further advocate for a pan-European resource that lists all interactions between varieties of different languages in multilingual continua. We propose organizing those around L languages, i.e., languages with weak standardization. For each variety of L, the entry should list all interacting language varieties, grounding those of H languages in respective diaglossic hierarchies. Table 1 illustrates such interaction entries for multilingual continua from Figure 2.

Diaglossic labeling of speech corpora. Finally, and most importantly, better cross-variety transfer requires actual speech corpora to be labeled with diaglossic variety information *at the instance level*. As we will report in §3.3, speech samples in most

existing resources are, at best, diatopically labeled, providing the geographic origin of the speaker. Unless the speaker is explicitly instructed to speak at a specified diaglossic height (typically a basilect), in the general case, we cannot know the diaglossic level of the recorded speech (e.g., speech recorded in Buzet, Croatia may still better represent the Istrian regiolect or even standard Croatian than the Buzet basilect). Sometimes, the diaglossic level may be inferred, e.g., from demographic information: older and less educated speakers tend to speak lower varieties (Trudgill, 2000). Such inference, however, is noisy and usually possible only for basilects. We thus strongly advocate for annotating speech instances with vertically closest varieties of respective diaglossic hierarchies.

3.3 The Landscape of Speech Resources

Finally, we inspect a non-exhaustive but broad inventory of 86 speech resources of European varieties, focusing on resources likely to contain at least some dialectal speech; our aim is to determine the extent to which the diaglossic dimension is documented. We group resources based on geographic resolution into five categories: village/exact point, urban area/subregion, region, country, and supranational linguistic area. For each resource, we annotate the following: (i) *declared* diaglossic level: the predominant diaglossic level of speech present in the data, as indicated (if at all) by the creators of the resource; (ii) *estimated* diaglossic level: realizing that very few resources have a declared diaglossic level, we annotate what we believe to be the predominant diaglossic level of speech in the resource, based on all available information or even by listening to the speech samples; (iii) whether the speech is transcribed into text in the same language and the diaglossic level of the transcription (i.e., the written variety); and (iv) whether the resource includes ethnolect speech. Additionally, we indicate how available/accessible the actual speech data is.

Table 2 in the Appendix details all the examined resources; for brevity, here we discuss only the aggregate statistics. The examined resources display a fairly balanced distribution over the diatopic resolution: 20 resources record exact points of speaker origin (typically villages in rural areas), 16 denote urban areas/subregions, 31 indicate the larger region, and for 13 we only know the country. Unfortunately but expectedly, very few resources indicate the diaglossic variety of the recorded speech: the vast majority (75) have no diaglossic information, and of the few that do, most (5) indicate a basilect. Our effort to estimate diaglossic levels reveals why indicating just the diatopic resolution is not enough: out of 28 resources for which we estimated the level of *high standard*, for 17 of them, the speech samples were designated with sub-national diatopic origin (i.e., regions, urban/sub-region areas, or villages/exact points). Even if we assume some noise in our estimates of the diaglossic height, this strongly suggests that reliably inferring the extent of presence of dialectal/standard features (i.e., the diaglossic height of the speech) from diatopic information alone is not possible. Explicit annotation of diaglossic levels of speech is thus necessary for a more effective cross-variety transfer and, in turn, dialectally inclusive speech technology.

Encouragingly, most resources (79) have transcriptions, but most of them (32) are in high standards (14 have a basilectal text, 3 regiolectal; for 36 there is no indication of diaglossic level of transcription). Ten of the examined resources cover speech from ethnolectal varieties. Finally, and positively, most resources (67) are available to researchers, either directly (49) or upon request (18).

4 Conclusion

In this position paper, we examined the European linguistic landscape and analyzed the restrictive, unidimensional diatopic lens through which it is computationally modeled. A crucial, *diaglossic* dimension—the vertical continuum between the traditional dialects and the high standard of a language induced by standardization efforts—is largely ignored in both resource creation and computational modeling. We examine 86 speech datasets, uncovering the critical gap between the diatopic and diaglossic annotation: speech collected from geographically localized areas—with the aim of capturing "dialectal" features—is often much closer to the high standard than traditional dialect. Supporting dialectal varieties requires a much more effective cross-variety model transfer, for which both computational modeling of diaglossic continua and interactions between varieties are required. To this effect, we offer recommendations for: (i) creating (computer-readable) diaglossic hierarchies of modern standardized European languages, and (ii) interaction registers, documenting localized interactions between varieties of different languages, and, crucially, (iii) annotating dialectal speech with diaglossic distances to high-standard varieties.

Limitations

We recognize three main limitations of our work.

(1) We focus on and advocate for computational modeling of *diaglossia*, because of its pervasiveness in European linguistic landscape and impact on speech technology, but also because it is a dimension of variation in the diasystem that is, arguably, the easiest to formalize and capture. Our contribution is limited to diaglossia; we do not consider other dimensions of variation, namely diastratic, diaphasic, and diamesic effects. A complete (computational) treatment of linguistic variation would have to model these dimensions too, including their inevitable interaction with the diaglossic dimension of variation.

(2) Both in terms of the analyzed geographic areas and examined speech resources, we limit ourselves to variation of European languages *on the European continent*; many European languages are, as a consequence of colonization and imperialism, spoken on other continents too, where they have developed interesting varieties, often due to interaction with indigenous languages; these varieties, however, are outside of the scope of our analysis.

(3) Finally, we examine a significant, but far from exhaustive set of speech resources. While we believe that the sample is large enough to draw conclusions about the extent of diaglossic information in speech resources more generally (and produce corresponding recommendations), it is conceivable that a more exhaustive investigation would yield somewhat deviating findings.

Ethics Statement

Our work advocates for a linguistically grounded and more computationally efficient modeling of European linguistic variation in LT. Our quantification of the dialectal speech technology gap focuses on ASR for basilects, ethnolects, and varieties of (closely) related local languages L in different contexts; however, whether building ASR models—and creating written transcriptions—is appropriate for a specific language variety should be assessed on a case-by-case basis, considering functional differentiation of co-existing languages in multilingual spaces. This is especially crucial when dealing with dialects of small, local, endangered languages of Europe, for which direct engagement with speech communities is not only necessary to avoid extractive research practices and cultural appropriation (Ògúnremí et al., 2023; Bird, 2024), but also central to understand local needs to avoid misrepresenting minority languages and cultures (Ramponi, 2024).

Acknowledgment of AI Use.

We used Claude Opus 4.7 to convert the results of our resource survey, recorded in an Excel sheet, into a longtable Latex format.

References

Noëmi Aepli and Rico Sennrich. 2022. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2024. MEGAVERSE: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637, Mexico City, Mexico. Association for Computational Linguistics.

Tuka Alhanai, Adam Kasumovic, Mohammad M. Ghassemi, Aven Zitzelberger, Jessica M. Lundin, and Guillaume Chabot-Couture. 2025. Bridging the gap: Enhancing LLM performance for low-resource African languages with new benchmarks, fine-tuning, and cultural adjustments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27802–27812.

Peter Auer. 2005. Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations. In *Perspectives on Variation: Sociolinguistic, Historical, Comparative*, pages 7–42. De Gruyter Mouton.

Peter Auer. 2011. Dialect vs. standard: a typology of scenarios in Europe. In *The Languages and Linguistics of Europe – A Comprehensive Guide*, The World of Linguistics, pages 485–500. De Gruyter.

Peter Auer. 2017. Dialect change in Europe—leveling and convergence. *The Handbook of Dialectology*, pages 159–176.

Gotzon Aurrekoetxea, Ariane Ensunza, Jožica Škofic, and Hans Van de Velde. 2022. An introduction to dialect classifications in Europe. *Dialectologia*, page 1–17.

Sjef Barbiers, Leonie Cornips, and Jan Pieter Kunst. 2007. The syntactic atlas of the Dutch dialects (SAND): A corpus of elicited speech and text as an

online dynamic atlas. In Joan C. Beal, Karen P. Corrigan, and Hermann L. Moisl, editors, *Creating and Digitizing Language Corpora*, volume 1: Synchronic Databases, pages 54–90. Palgrave Macmillan UK, London.

Mario Barcala, Eva Domínguez, Alba Fernández, Raquel Rivas, Maria Paula Santalla, Victoria Vázquez, and Rebeca Villapol. 2018. El corpus ESLORA de español oral: diseño, desarrollo y explotación. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, 5(2):217–237.

Steven Bird. 2024. Must NLP be extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14915–14929, Bangkok, Thailand. Association for Computational Linguistics.

Verena Blaschke, Christoph Purschke, Hinrich Schuetze, and Barbara Plank. 2024. What do dialect speakers want? A survey of attitudes towards language technology for German dialects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 823–841, Bangkok, Thailand. Association for Computational Linguistics.

Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023a. A survey of corpora for Germanic low-resource languages and dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.

Verena Blaschke, Hinrich Schütze, and Barbara Plank. 2023b. Does manipulating tokenization aid cross-lingual transfer? A study on POS tagging for non-standardized languages. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 40–54, Dubrovnik, Croatia. Association for Computational Linguistics.

Verena Blaschke, Miriam Winkler, Constantin Förster, Gabriele Wenger-Glemser, and Barbara Plank. 2025. A multi-dialectal dataset for German dialect ASR and dialect-to-standard speech translation. In *Proc. Interspeech 2025*, pages 913–917. ISCA.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Michael Cacioli, Liam Eggleston, Jatin Sarabu, and Kevin Zhu. 2026. Can you hear Naples? Building and benchmarking a Neapolitan speech corpus. In *Proceedings of the AAAI 2026 Workshop on Audio-Centric AI: Towards Real-World Multimodal Reasoning and Application Use Cases (Audio-AAAI)*, volume 312 of *Proceedings of Machine Learning Research*, pages 94–112. PMLR.

A. Castelruiz, J. Sanchez, X. Zalbide, E. Navas, and I. Graminde. 2004. Description and design of a web accessible multimedia archive. In *Proceedings of the 12th IEEE Mediterranean Electrotechnical Conference (IEEE Cat. No.04CH37521)*, volume 2, pages 681–684.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024. Towards robust speech representation learning for thousands of languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10205–10224, Miami, Florida, USA. Association for Computational Linguistics.

Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. Association for Computational Linguistics.

Grzegorz Chrupała. 2023. Putting natural in natural language processing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7820–7827, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Ankur Bapna, Yu Zhang, Min Ma, Patrick von Platen, Anton Lozhkov, Colin Cherry, Ye Jia, Clara Rivera, Mihir Kale, and 1 others. 2022. XTREME-S: Evaluating cross-lingual speech representations. In *Proc. Interspeech 2022*, pages 3248–3252.

Eugenio Coseriu. 1956. *La geografía lingüística*. Instituto de Filología, Facultad de Humanidades y Ciencias, Universidad de la República, Montevideo, Uruguay.

Carlota de Benito Moreno, Javier Pueyo, and Inés Fernández-Ordóñez. 2016. Creating and designing a corpus of rural Spanish. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 78–83.

Ferdinand de Saussure. 1916. *Cours de linguistique générale*, 1st edition. Payot, Lausanne, Switzerland and Paris, France.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *The World Atlas of Language Structures Online (v2020.4)*. Zenodo.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. DIALECTBENCH: An nlp benchmark for dialects, varieties, and closely-related languages. In *Proceedings of the 62nd*

Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 14412–14454, Bangkok, Thailand. Association for Computational Linguistics.

Tiantian Feng, Kevin Huang, Anfeng Xu, Xuan Shi, Thanathai Lertpetchpun, Jihwan Lee, Yoonjeong Lee, Dani Byrd, and Shrikanth Narayanan. 2026. Voxlect: A speech foundation model benchmark for modeling dialects and regional languages around the globe. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 2640–2651.

J.A. Fishman. 2001. Why is it so hard to save a threatened language? In Joshua A. Fishman, editor, *Can Threatened Languages be Saved?*, pages 1–22. Multilingual Matters, Bristol, Blue Ridge Summit.

Elisabeth Fittschen, Sabrina Xin Li, Tom Lippincott, Leshem Choshen, and Craig Messner. 2026. Pretraining language models for diachronic linguistic change discovery. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4627–4642, Rabat, Morocco. Association for Computational Linguistics.

Hans Goebl. 1986. Problems and perspectives of contact linguistics from a Romance scholar's point of view. In *Language contact in Europe. Proceedings of the Working Croups 12 and 13 at the XIIIth International Congress of Linguists, August 29—September 4, 1982, Tokyo*, pages 125–150. Max Niemeyer Verlag.

A. Goeman and J. Taeldeman. 1996. Fonologie en morfologie van de Nederlandse dialecten: Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. *Taal en Tongval*, 48:38–59.

Matthew J. Gordon. 2017. Structural dialectology. *The Handbook of Dialectology*, pages 73–87.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2026. Glottolog 5.3. Leipzig: Max Planck Institute for Evolutionary Anthropology. Available online at http://glottolog.org.

Steffen Höder. 2014. Constructing diasystems. In Tor A. Åfali and Brit Mæhlum, editors, *The Sociolinguistics of Grammar*, pages 137–152. John Benjamins.

Valentin Hofmann, Goran Glavaš, Nikola Ljubešić, Janet B. Pierrehumbert, and Hinrich Schütze. 2024. Geographic adaptation of pretrained language models. *Transactions of the Association for Computational Linguistics*, 12:411–431.

Ingrid Hove. 2013. Prosodic differences between Germans and German-speaking Swiss in L2. *Revue Tranel (Travaux neuchâtelois de linguistique)*, 59:57–70.

Péter Jeszenszky and Robert Weibel. 2015. Measuring boundaries in the dialect continuum. In *18th AGILE International Conference on Geographic Information Science, Lisbon (Portugal)*.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2025. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, 57(6):1–37.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Roland Kehrein. 2019. Vertical language change in Germany: Dialects, regiolects, and standard German. In Stanley D. Brunn and Roland Kehrein, editors, *Handbook of the Changing World Language Map*. Springer International Publishing.

Lennart Keller and Goran Glavaš. 2025. SpeechTaxi: On multilingual semantic speech classification. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pages 1–5. IEEE.

Heinz Kloss. 1967. 'Abstand languages' and 'ausbau languages'. *Anthropological Linguistics*, 9(7):29–41.

Heinz Kloss. 1978. *Die Entwicklung neuer germanischer Kultursprachen seit 1800*, 2nd edition. Sprache der Gegenwart. Schwann.

Zuzana Komrsková, Marie Kopřivová, David Lukeš, Petra Poukarová, and Hana Goláňová. 2017. New spoken corpora of Czech: ORTOFON and DIALEKT. *Jazykovedny Casopis*, 68(2):219–228.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1384–1397, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8186–8213, Bangkok, Thailand. Association for Computational Linguistics.

Juliane Limper, Jeffrey Pheiff, and Anneli Williams. 2019. REDE SprachGIS: A geographic information system for linguists. In *Handbook of the Changing World Language Map*, pages 3743–3771. Springer.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Anne-Marie Lutgen, Alistair Plum, Verena Blaschke, Barbara Plank, and Christoph Purschke. 2026. Variation is the norm: Embracing sociolinguistics in NLP. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 11332–11344, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Malak Mashabi, Shahad Al-Khalifa, and Hend Al-Khalifa. 2024. A survey of large language models for Arabic language and its dialects. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Alberto M. Mioni. 1983. Italiano tendenziale: osservazioni su alcuni aspetti della standardizzazione. In *Scritti linguistici in onore di Giovan Battista Pellegrini*, volume 1, pages 495–517, Pisa, Italy. Pacini.

Steven Moran and Daniel McCloy. 2019. PHOIBLE 2.0. Jena: Max Planck Institute for the Science of Human History.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv*, 2207.04672.

Tolúlope Ògúnremí, Wilhelmina Onyothi Nekoto, and Saron Samuel. 2023. Decolonizing NLP for "low-resource languages": Applying Abebe Birhane's relational ethics. *GRACE: Global review of AI community Ethics*, 1(1).

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *arXiv*, 2511.09690.

Barbara Plank. 2016. What to do about non-standard (or non-canonical) language in NLP. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 13–20.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv*, 2212.04356.

Alan Ramponi. 2024. Language varieties of Italy: Technology challenges and opportunities. *Transactions of the Association for Computational Linguistics*, 12:19–38.

Fabian David Schmidt, Ivan Vulić, Goran Glavaš, and David Ifeoluwa Adelani. 2025. Fleurs-SLU: A massively multilingual benchmark for spoken language understanding. In *Second Conference on Language Modeling*.

Jürgen Erich Schmidt. 2011. Formation of and change in regiolects and (regional) dialects in German. *Taal en Tongval*, 63(1):143–173.

Christoforos Souganidis, Asier Herranz, Ibon Saratxaga, Eva Navas, and Inma Hernáez. 2026. Phonologically-aware automatic speech recognition evaluation of low-resource languages: The case of Basque dialects. In *Proceedings of DialRes Workshop at LREC26*.

Jane Stuart-Smith, Morgan Sonderegger, Jeff Mielke, James Tanner, Vanna Willerton, and Rachel Macdonald. 2024. SPADE – Speech across dialects of English. OSF. Accessed: 2026-05-01.

Peter Trudgill. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in society*, 3(2):215–246.

Peter Trudgill. 1986. *Dialects in Contact*, volume 10 of *Language in Society*. Blackwell.

Peter Trudgill. 1992. The Ausbau sociolinguistics of Greek as a minority and majority language. *Selected papers on theoretical and applied linguistics*, 6:213–235.

Peter Trudgill. 2000. *Sociolinguistics: An Introduction to Language and Society*. Penguin UK.

Jo Verhoeven. 2005. Belgian Standard Dutch. *Journal of the International Phonetic Association*, 35(2):243–247.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Uriel Weinreich. 1954. Is a structural dialectology possible? *Word*, 10(2-3):388–400.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *arXiv*, 2504.15521.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Marcely Zanon Boito, Antonios Anastasopoulos, Aline Villavicencio, Laurent Besacier, and Marika Lekakou. 2018. A small Griko-Italian speech translation corpus. In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 36–41.

A Appendix

Table 2: A non-exhaustive overview of (dialectal) speech corpora of European languages.

Resource	Standard language	Diatopic resolution	Declared diaglossic level	Estimated diaglossic level	Transcr.?	Diaglossic level of transcrip-tion	Ethnolect?	Access
SwissGPC 1.0	German	Country	No info	Regiolect	✗	No info	✗	Open
Common Voice Spanish Read Speech	Spanish	Supranational linguistic area	No info	High standard	✓	High standard	✗	Open
Common Voice Spanish Spontaneous Speech	Spanish	Supranational linguistic area	No info	Regiolect	✓	High standard	✗	Open
Common Voice Aragonese Read Speech	Aragonese	Region	No info	High standard	✓	High standard	✗	Open
Common Voice Aragonese Spontaneous Speech	Aragonese	Region	No info	Regiolect	✓	High standard	✗	Open
Common Voice Asturian Read Speech	Asturian	Region	No info	High standard	✓	High standard	✗	Open
Common Voice Basque Read Speech	Basque	Urban area / subregion	No info	High standard	✓	High standard	✗	Open
Common Voice Catalan Read Speech	Catalan	Urban area / subregion	No info	High standard	✓	High standard	✗	Open
Common Voice Catalan Spontaneous Speech	Catalan	Urban area / subregion	No info	Regiolect	✓	High standard	✗	Open
Common Voice Galician Read Speech	Galician	Urban area / subregion	No info	High standard	✓	High standard	✗	Open
Common Voice Galician Spontaneous Speech	Galician	Region	No info	Regiolect	✓	High standard	✗	Open
Common Voice French Read Speech	French	Supranational linguistic area	No info	High standard	✓	High standard	✗	Open
Common Voice French Spontaneous Speech	French	Supranational linguistic area	No info	Regiolect	✓	High standard	✗	Open
Common Voice Breton Read Speech	Breton	Urban area / subregion	No info	High standard	✓	High standard	✗	Open
Common Voice Breton Spontaneous Speech	Breton	Region	No info	Regiolect	✓	High standard	✗	Open
Common Voice Occitan Read Speech	Occitan	Region	No info	High standard	✓	High standard	✗	Open
Common Voice Alsatian Read Speech	German	Urban area / subregion	No info	High standard	✓	High standard	✗	Open

Resource	Standard language	Diatopic resolution	Declared diaglossic level	Estimated diaglossic level	Transcr.?	Diaglossic level of transcrip-tion	Ethnolect?	Access
Common Voice Alsatian Spontaneous Speech	German	Region	No info	Regiolect	✓	High standard	✗	Open
Common Voice Portuguese Read Speech	Portuguese	Supranational linguistic area	No info	High standard	✓	High standard	✗	Open
COSER – Corpus Oral y Sonoro del Español Rural	Spanish	Village / exact point	No info	Basilect	✓	No info	✗	On Request
PRESEEA – Proyecto para el Estudio Sociolingüístico del Español de España y América	Spanish	Village / exact point	No info	High standard	✓	No info	✗	Open
EDA – Euskararen Herri Hizkeren Atlasa	Basque	Village / exact point	No info	Basilect	✓	No info	✗	Open
Euskal Herriko hizkerak eta ahozko ondarea	Basque	Village / exact point	No info	Basilect	✓	No info	✗	Unavailable
Bizkaieraren Bideoteka	Basque	Village / exact point	No info	Basilect	✓	Basilect	✗	Unavailable
AGO – Arquivo do Galego Oral	Galician	Village / exact point	No info	Basilect	✓	High standard	✗	On Request
Nos_ParlaSpeech-GL: Galician ASR corpus	Galician	Region	No info	High standard	✓	No info	✗	On Request
CORILGA – Corpus Oral Informatizado da Lingua Galega	Galician	Region	No info	Basilect	✓	No info	✗	On Request
Las lenguas de Aragón	Aragonese	Village / exact point	No info	Basilect	✓	No info	✗	Unavailable
CFPP2000 – Corpus de Français Parlé à Paris	French	Village / exact point	No info	High standard	✓	No info	✗	Open
Portuguese SpeechDat(M)	Portuguese	Country	No info	High standard	✓	No info	✗	Paid
Portuguese SpeechDat(II) FDB-4000	Portuguese	Country	No info	High standard	✓	No info	✗	Paid
Basque FDB-1060 database	Basque	Region	No info	High standard	✓	No info	✗	Paid
Catalan SpeechDat-Car database	Catalan	Region	No info	High standard	✓	No info	✗	Paid
Catalan-SpeechDat For the Fixed Telephone Network Database	Catalan	Region	No info	High standard	✓	No info	✗	Paid
SpeechDat Catalan FDB database	Catalan	Region	No info	High standard	✓	No info	✗	Paid
Catalan-SpeechDat for the Mobile Telephone Network Database	Catalan	Region	No info	High standard	✓	No info	✗	Paid
Spanish SpeechDat-Car database	Spanish	Country	No info	High standard	✓	No info	✗	Paid

Resource	Standard language	Diatopic resolution	Declared diaglossic level	Estimated diaglossic level	Transcr.?	Diaglossic level of transcrip-tion	Ethnolect?	Access
Spanish SpeechDat database for the mobile telephone network	Spanish	Country	No info	High standard	✓	No info	✗	Paid
Spanish SpeechDat(II) FDB-1000 database	Spanish	Country	No info	High standard	✓	No info	✗	Paid
Spanish SpeechDat(II) FDB-4000	Spanish	Country	No info	High standard	✓	No info	✗	Paid
SpeechDat(M) Spanish database	Spanish	Country	No info	High standard	✓	No info	✗	Paid
SpeechDat Galician Database for the Fixed Telephone Network	Galician	Region	No info	High standard	✓	No info	✗	Paid
MEDIA speech database for French	French	Country	No info	No info	✓	No info	✗	Paid
CORAL Corpus	Portuguese	Urban area / subregion	No info	Urban/sub-regional dialect	✓	No info	✗	Paid
Fundamental Portuguese Corpus	Portuguese	Country	No info	No info	✓	No info	✗	Paid
LECTRA (LECture TRAnscriptions in European Portuguese)	Portuguese	Country	No info	High standard	✓	No info	✗	Paid
CORDIAL-SIN	Portuguese	Village / exact point	No info	Basilect	✓	No info	✗	Open
VIVALDI	Many (from Romance, Germanic, Slavic, Hellenic, and Albanian families)	Village / exact point	No info	Basilect	✗	No info	✓	Unavailable
KIParla	Italian	Region	Regiolect	Regiolect	✓	Regiolect	✗	On Request
ITALIC	Italian	Region	Regional standard	Regional standard	✓	High standard	✗	Open
UOI	Italiot-Greek / Modern Greek	Village / exact point	No info	Basilect	✓	Basilect	✓	Open
Na-našu (Acquaviva)	Molise Slavic	Village / exact point	No info	Basilect	✓	Basilect	✓	Open
Na-našu (Montemitro)	Molise Slavic	Village / exact point	No info	Basilect	✓	Basilect	✓	Open
Na-našu (San Felice)	Molise Slavic	Village / exact point	No info	Basilect	✓	Basilect	✓	Open
Common Voice Spontaneous Speech 3.0 - Ligurian	Ligurian	Region	No info	Basilect	✓	Basilect	✗	Open
Common Voice Scripted Speech 25.0 - Ligurian	Ligurian	Region	No info	Regional standard	✓	High standard	✗	Open

Resource	Standard language	Diatopic resolution	Declared diaglossic level	Estimated diaglossic level	Transcr.?	Diaglossic level of transcrip-tion	Ethnolect?	Access
Common Voice Scripted Speech 25.0 - Sardinian	Sardinian	Region	No info	Regional standard	✓	High standard	✗	Open
Common Voice Scripted Speech 25.0 - Romansh Vallader	Romansh	Region	No info	Regional standard	✓	High standard	✗	Open
AlpiLinK Corpus 1.2.1	Many (from Romance, Germanic, and Slavic families)	Village / exact point	Basilect	Basilect	✗	No info	✓	On Request
Neapolitan Spoken Corpus	Neapolitan	Urban area / subregion	Urban/sub-regional dialect	Urban/sub-regional dialect	✓	Urban/sub-regional dialect	✗	Open
GOS: Corpus of Spoken Slovene	Slovenian	Urban area / subregion	No info	No info	✓	High standard	✗	Open
Spoken Torlak Corpus	Serbian	Region	No info	Regiolect, Regional standard	✓	Regional standard	✗	Open
HrAL (part ot TalkBank)	Croatian	Country	No info	No info	✓	Regiolect	✗	Open
DIALEKT v1	Czech	Region	No info	Regiolect	✓	Regiolect, High standard	✗	On Request
Corpus ASD	German (source roof: German, target roof: Roma-nian)	Region	No info	Regiolect	✓	Basilect	✓	On Request
Estonian Dialectal Corpus	Estonian	Region	No info	No info	✓	High standard	✗	On Request
RoDia	Romanian	Region	No info	No info	✗	No info	✗	Open
Follow-up Study of Dialects of Finnish	Finnish	Urban area / subregion	No info	Urban/sub-regional dialect	✓	High standard	✗	On Request
The Finnish Dialect Syntax Archive	Finnish	Village / exact point	No info	Urban/sub-regional dialect, No info	✓	No info	✗	Open
Raja-Karjalan korpus	Finnish (Karelian is also considered a language of its own)	Village / exact point	No info	No info	✓	Basilect	✗	Open
LIA corpus	Norwegian	Village / exact point	No info	Basilect	✓	High standard	✗	Open
Nordic Dialect Corpus v. 4.0	Norwegian, Swedish, Danish, Faroese, Icelandic, Övdalian	Urban area / subregion	No info	Regiolect	✓	High standard	✗	No info
Common Voice Read Speech - Aromanian	Romanian	Supranational linguistic area	No info	No info	✓	No info	✓	Open
Common Voice Read Speech - Romanian	Romanian	Region	No info	Regiolect	✓	High standard	✗	Open

Resource	Standard language	Diatopic resolution	Declared diaglossic level	Estimated diaglossic level	Transcr.?	Diaglossic level of transcrip-tion	Ethnolect?	Access
Phonogrammarchiv - Ph 103	Romanian	Village / exact point	No info	Basilect	✗	No info	✓	On Request
Corpus of Slavic Dialects in Albania	Macedonian Serbo-Croatian	Urban area / subregion	Urban/sub-regional dialect	Urban/sub-regional dialect	✓	High standard, Regional standard	✓	No info
Corpus RVG1_CLARIN	German	Urban area / subregion	Urban/sub-regional dialect	Urban/sub-regional dialect	✓	High standard	✗	No info
Common Voice Spontaneous Speech 3.0 - Gheg Albanian	Albanian	Region	No info	No info	✓	High standard	✗	Open
Mintzoak	Basque	Region	Basilect, No info	Basilect	✓	Basilect	✗	Open
Euskalkiak	Basque	Country	No info	Urban/sub-regional dialect	✗	Basilect	✗	Open
xSID-audio	German	Urban area / subregion	No info	Urban/sub-regional dialect	✓	Basilect	✗	On Request
Betthupferl	German	Urban area / subregion	No info	Urban/sub-regional dialect	✓	Basilect, High standard	✗	Open
Korpus Sprachvariation in Norddeutschland (SiN)	German	Village / exact point	Basilect	Basilect	✗	No info	✗	On Request
NB Tale – Speech Database for Norwegian	Norwegian	Urban area / subregion	Basilect	Basilect	✓	High standard	✗	Open
SwissDial	German	Region	Basilect	Basilect	✓	Basilect, High standard	✗	On Request
walliserdeutsch/rro	German	Region	Basilect	Basilect	✓	Basilect, High standard	✗	On Request